NCBI BlastRules Collection Release Notes

Version 1.1, November 27, 2017

BlastRules from NCBI is a set of a set of manually created protein family classifiers, with attached metadata such as recommended protein product name. It is designed to work together with other sources of evidence, such as protein profile hidden Markov models (HMMs), so automated prokaryotic genome annotation pipelines can provide the possible functional annotation. BlastRules eventually will have a publicly accessible web page, but does not have one yet.

The required elements of each BlastRule are an **accession** number, a **rule type**, a descriptive **protein name**, a list of one or more reference **proteins**, and a set of **threshold** values that help determine if a BlastRule hits a protein. Optional elements include a gene symbol, one or more PubMed identifiers, and explanatory text. Future releases will include Enzyme Commission (EC) numbers. Rules may have additional information that is not provided in the current release, including tracking information such as owner and creation date.

This release of BlastRules does not include software for comparing target proteins to BlastRule reference proteins in order to determine whether or not a BLAST search result should be interpreted as a hit to a rule. Software that performs such evaluation is included in NCBI's Prokaryotic Genome Annotation Pipeline (PGAP), but PGAP itself is not currently publicly available.

Release Schedule: The release schedule has not been determined. The tentative plan is to provide quarterly releases. Release 1.1 is provided to coincide with electronic advance publication of Haft, et al., *RefSeq: an update on prokaryotic genome annotation and curation*, for the Nucleic Acids Res. annual database issue. Print publication is scheduled January of 2018. The article can be found at:
https://academic.oup.com/nar/article/doi/10.1093/nar/gkx1068/4588110?guestAccessKey=f2ba9d85-8124-4d93-80dc-45667aeab71b

Release Statistics:

Release	Date	Rules	Chosen evidence	Total evidence
1.0	2017-09-15	840	63,644 proteins	162,890 proteins
1.1	2017-12-08	5568	120,786 proteins	222,497 proteins

Note that coverage of RefSeq proteins may be understated for some releases of NCBI BlastRules because of the time required to process, activate, and apply newly created rules.

Content and Coverage: BlastRules currently serve primarily as a mechanism to address limits in the expressive power of available annotation rules from other sources, such as CDD-SPARCLE architectures or *equivalog* HMMs. Most BlastRules so far, therefore, cover narrowly focused,

high-interest, low-abundance proteins such as lineage-specific virulence proteins, serotype-specific markers, or vaccine candidate surface proteins from major human pathogens. Future waves of BlastRule creation, however, may address different sets of proteins and may change the overall character of database content and breadth of coverage by annotation pipelines.

The content of the BlastRules database is provided by appropriately named columns in a tab-separated values file. The columns are as follows:

BlastRule_Acc: BlastRule accessions take the form NBRnnnnnn, where 'n' is any digit.

Rule type: BlastRules compete with each other, and with other types of evidence, for the right to determine the name a protein should receive. The rule type determines precedence, on an arbitrary scale. So far, three types of rules are defined. The rule type also determines default values for three threshold parameters. *Identity* is the threshold for the percent identity between a BlastRule reference protein and candidate matching sequence, ignoring gap regions. *Model coverage* is the percent of the length of the reference sequence that must be aligned for a hit to be recognized. *Target coverage* is the percent length of the sequence being tested for a match to the rule. Current default values are 94% identity, 90% model coverage, 90% target coverage for *BlastRuleException*, and 80% identity, 90% model coverage, 80% target coverage, for *BlastRuleEquivalog*. However, the curator of a rule can change these thresholds as needed.

Note that the term "*equivalog*" describes a set of proteins that share a specific function by virtue of evolutionary descent from an ancestral sequence with that same function. *BlastRuleException* rules attach names that are even more specific than *equivalog* names, as when the literature distinguishes among different isozymes from the same *equivalog* family, or among closely related virulence proteins.

Name: Names from BlastRules are designed to comply with the standards that GenBank and RefSeq require for valid protein names, and to become the protein product name in PGAP and RefSeq annotations. A typical name, "tandem repeat protein effector TRP47", follows the literature as closely as it can, but avoids a discouraged and troublesome explicit reference to molecular weight, "47 kDa." Proteins recognized by the rule, in fact, are repeat-rich and quite variable in length, and but are properly named "TRP47" irrespective of the computed molecular mass.

Proteins: Proteins used as reference sequences for BlastRules are given as a comma-separated list of protein accessions. It is better to use several sequences for a BlastRule, and fairly high cutoffs (all proteins in a rule use the same cutoffs) than to use a single sequence only and overly permissive cutoffs. The set of proteins used to define a BlastRule may be considered a working list for development of an HMM in the future. All proteins are treated as full length; care should be taken to use only sequences with correct start sites while building rules.

Identity: A BlastRule hits a target protein if three conditions are met: percent identity, percent of the BlastRule's model protein aligned to the model by BLAST, and percent of the target protein's length aligned to the model. The first of these is Identity, typically 94% for *BlastRuleException*, 80% for *BlastRuleEquivalog.*

Model_pct: The protein(s), used to define a BlastRule is(are) the model(s) for that rule. Model_pct is simply the miminum percent of a model protein's length that must be used by BLAST to align to a target protein that the BlastRule identifies as a match. Post-processing of BLAST alignments is presumed, so two or more hit regions can be combined as long as no region of either the model protein or the query protein can be used twice and aligned regions are in the same order on both proteins.

Target_pct: This number is the the miminum percent of a target protein's length that must match to a BlastRule model protein for evidence of a BlastRule hit to register. It measures on a target protein what Model_pct measures on a BlastRule model protein. Note that we refer to Model and Target, rather than Query and Subject, because the choice of which sequences to treat as the query and which to make subjects in a BLAST-searchable database may not always be obvious.

Gene: This optional field must have a single value only that complies with standards for GenBank's standards for gene symbols in prokaryotic genomes, or else is left null.

PMID: PubMed identifiers describe a BlastRule on the whole, although users may be able to infer how individual PMID link to individual model proteins.

Comment: Each rule may have a plain text comment consisting of explanatory text that eventually could accompany the public annotation of a protein, as through the Entrez query system. The comment must be formatted as a single line (no carriage returns or line feeds allowed).

Attribution: Rule authorship is tracked. A single rule may have multiple authors. BlastRule distribution files will include attribution in future releases once rules developed in collaboration with outside experts are included. In the current release, all BlastRules were developed by NCBI curators. Thousands of rules were derived, within NCBI, from expertly curated data on insertion sequence transposases compiled by ISFinder – see Siguier, et al., *Exploring bacterial insertion sequences with ISfinder: objectives, uses, and future developments*, Methods Mol Biol. 2012;859:91-103 (PMID:22367867).